

Missing Data and the EM Algorithm

Guy Lebanon

December 13, 2006

A maximum likelihood estimator $\hat{\theta} = \arg \max_{\theta} \ell(\theta)$ for observed data is often difficult to compute if the data model is specified in terms of a joint over observed and missing data. In other words, the data model $p_{\theta}(x, z)$ describes the generation of the observed data x , but also of unobserved data z . A maximum likelihood approach would maximize the probability of the observed data

$$\hat{\theta}_{\text{mle}} = \arg \max_{\theta} \ell(\theta) = \arg \max_{\theta} \log \prod_{i=1}^n p(x^{(i)}) = \arg \max_{\theta} \sum_{i=1}^n \log \sum_{z^{(i)}} p_{\theta}(x^{(i)}, z^{(i)}). \quad (1)$$

The problem with (1) is that the inner summation over all possible values of the missing data $z^{(i)}$ may be difficult to compute. For some models $p_{\theta}(x, y)$ it may grow exponentially (for example hidden Markov models) and computing the gradient and likelihood for use in a gradient ascent maximization is infeasible. The EM algorithm is an iterative algorithm that solves (1) without the need to perform the inner summation in (1) (at least for some models). Using Jensen's inequality it constructs a lower bound on the likelihood and instead of maximizing $\ell(\theta)$, it maximizes the bound.

Proposition 1 (Jensen, [1]). *For a RV X and a convex function f we have $E f(X) \geq f(EX)$. Moreover, if f is strictly convex, equality holds iff X is degenerate i.e. $P(X = EX) = 1$.*

The EM algorithm is based on maximizing the following bound on the likelihood of the observed data

$$\ell(\theta) = \sum_{i=1}^n \log \sum_{z^{(i)}} p_{\theta}(x^{(i)}, z^{(i)}) = \sum_{i=1}^n \log \sum_{z^{(i)}} q_i(z^{(i)}) \frac{p_{\theta}(x^{(i)}, z^{(i)})}{q_i(z^{(i)})} = \sum_{i=1}^n \log E_{q_i} \left(\frac{p_{\theta}(x^{(i)}, z^{(i)})}{q_i(z^{(i)})} \right) \quad (2)$$

$$\geq \sum_{i=1}^n E \left(\log \frac{p_{\theta}(x^{(i)}, z^{(i)})}{q_i(z^{(i)})} \right) = \sum_{i=1}^n \sum_{z^{(i)}} q_i(z^{(i)}) \log \frac{p_{\theta}(x^{(i)}, z^{(i)})}{q_i(z^{(i)})} \quad (3)$$

where we use the fact that $-\log(x)$ is convex and we assume that q_i are non-zero probability distributions. Note that the denominator does not depend on θ and therefore can be removed in maximization over θ . Above, we actually have a parameterized family of bounds - one bound for each selection of the distributions q_1, \dots, q_n . Recall that Jensen's inequality is equality for deterministic RV and therefore the selection

$$q_i(z^{(i)}) \propto p_{\theta'}(x^{(i)}, z^{(i)}) \quad \Rightarrow \quad q_i(z^{(i)}) = \frac{p_{\theta'}(x^{(i)}, z^{(i)})}{\sum_{z^{(i)}} p_{\theta'}(x^{(i)}, z^{(i)})} = p_{\theta'}(z^{(i)} | x^{(i)})$$

would yield a tight bound with equality at $\theta = \theta'$.

The resulting algorithm of bound computation and maximization is called expectation maximization (EM)

E step: compute the bound on the observed likelihood

$$\begin{aligned}
Q(\theta, \theta^{(t)}) &\stackrel{\text{def}}{=} \sum_{i=1}^n \sum_{z^{(i)}} p_{\theta^{(t)}}(z^{(i)}|x^{(i)}) \log p_{\theta}(x^{(i)}, z^{(i)}) \\
&= \sum_{i=1}^n \sum_{z^{(1)}, \dots, z^{(n)}} p_{\theta^{(t)}}(z^{(1)}, \dots, z^{(n)}|x^{(1)}, \dots, x^{(n)}) \log p_{\theta}(x^{(i)}, z^{(i)}) \\
&= \sum_{z^{(1)}, \dots, z^{(n)}} p_{\theta^{(t)}}(z^{(1)}, \dots, z^{(n)}|x^{(1)}, \dots, x^{(n)}) \log p_{\theta}(x^{(1)}, z^{(1)}, \dots, x^{(n)}, z^{(n)}) \\
&= \mathbb{E} \left(\log p_{\theta}(x^{(1)}, z^{(1)}, \dots, x^{(n)}, z^{(n)}) \mid x^{(1)}, \dots, x^{(n)}, \theta^{(t)} \right)
\end{aligned}$$

M step: maximize the bound to obtain

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta, \theta^{(t)})$$

The fact that each iteration in the EM algorithm increases the likelihood may be seen by

$$\ell(\theta^{(t+1)}) \geq \sum_{i=1}^n \sum_{z^{(i)}} p_{\theta^{(t)}}(z^{(i)}|x^{(i)}) \log p_{\theta^{(t+1)}}(x^{(i)}, z^{(i)}) \geq \sum_{i=1}^n \sum_{z^{(i)}} p_{\theta^{(t)}}(z^{(i)}|x^{(i)}) \log p_{\theta^{(t)}}(x^{(i)}, z^{(i)}) \geq \ell(\theta^{(t)})$$

where the first inequality follows from Jensen's inequality and the second from the maximization step in EM.

Two models where EM is commonly used are mixture of Gaussians and hidden Markov models. We describe hidden Markov model in a future note. In mixture of Gaussians we have $X \in \mathbb{R}$ and $Z \in \{1, \dots, k\}$ and $p(X = x, Z = i) = p(X = x|Z = i)p(Z = i)$ where $p(X = x|Z = i)$ is Gaussian $N(\mu_i, \Sigma_i)$ and $p(Z = i) = \pi_i$. The Z variable is missing and a maximum likelihood estimator for $\theta = (\pi, \mu, \Sigma)$ based on the observed data is

$$\hat{\theta} = \arg \max_{\pi, \mu, \Sigma} \sum_{i=1}^n \log \sum_{z^{(i)}} p(x^{(i)}|z^{(i)})p(z^{(i)}) = \arg \max_{\pi, \mu, \Sigma} \sum_{i=1}^n \log \sum_{z^{(i)}} p(\mu_{z^{(i)}}, \Sigma_{z^{(i)}})(x^{(i)})\pi_{z^{(i)}}.$$

The straightforward approach would be to perform gradient descent to find the optimal θ . An alternative solution based on EM would be to iteratively compute and maximize $Q(\theta, \theta^{(t)})$.

In some cases EM would be slower than gradient descent. In other cases, for example hidden Markov models, naive gradient descent is extremely slow and perhaps even impossible to compute. Then the EM algorithm achieves its full strength by enabling previously impossible statistical computing.

References

- [1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, second edition, 2005.